**Breaking in to the mainstream: demonstrating the value of Internet (and web) histories[[1]](#endnote-1)**

Jane Winters, School of Advanced Study, University of London

The realization of the Internet’s potential to connect not just computers but individuals, families, communities and nations – through the growth of the web – has transformed our lives over the last two decades. Our histories are increasingly both created and consumed online, for an audience of millions or for an audience of only one or two people. The ease with which it is possible to write and post information online, the speed with which one can react to news and contribute to ensuing debates, has dramatically altered – in scale and type – the group of people whom we might now describe as creators, publishers or authors. While some voices are, of course, excluded from this discourse,[[2]](#endnote-2) and others still are hidden or devalued, many more have taken advantage of new technologies and social forums publicly to share their lives, thoughts and beliefs.[[3]](#endnote-3)

Consequently, after little more than a quarter of a century, the web already constitutes an unprecedentedly rich primary source, combining information from personal blogs, to formal reportage, to the communications of local and national government. It is where we socialize, learn, campaign and shop. All human life, as it were, is vigorously there. And while the live web is extraordinarily ephemeral, much of this data – an ever more important historical record – persists in web archives (Jackson, 2015).[[4]](#endnote-4) It might seem obvious that one could not begin to write the history of the late 20th and early 21st centuries without recourse to web archives. Yet, researchers in the humanities, and historians in particular, have proven reluctant to engage with the histories of the web and the Internet, to explore web archives, to learn how to work with a new kind of source.

There are many practical reasons for this historical blind spot, which are more or less difficult to address. The most significant barrier to working with web archives is, quite simply, that it is difficult; it requires skills that many historians do not have, and in the short term may be unwilling to learn; it involves acknowledging a degree of ignorance with which otherwise seasoned researchers may be uncomfortable. The various methodological and theoretical challenges have been well enumerated, even if we are far from having solutions for them all (see, e.g., Brügger, 2013; Brügger and Finnemann, 2013). The multiplicity of national legal frameworks which regulate access to and reuse of web archives are a considerable hurdle to even the most dedicated scholar, but can seem an insurmountable problem to the more casual user. In the UK, for example, legal deposit legislation guarantees a comprehensive crawl of the .uk domain at least once a year, but it also ensures that anyone interested in studying the results of that crawl has to go to a reading room in one of the six UK copyright libraries and work through the archive web page by web page. The Internet Archive’s Wayback Machine (<http://archive.org/web/>) does not have the same restrictions on access, and now supports searching on the text of archived home pages rather than simply by URL, but reuse remains a problem and rich analytical access is still a very distant prospect.[[5]](#endnote-5)

There is no doubt that these are real issues, but there is something else at work too. It is not the case that large numbers of historians are trying to work with web archives, struggling and then giving up. Lack of awareness is a far more important criterion. Researchers in the humanities are used to working with digitized historical sources that they access online, and to reading much of their secondary literature on the web too, but they do not seem ready for the archived web itself as a focus of research, as embodying important histories. There are some indications that this is beginning to change. In February 2016, for example, the Mile End Institute (Queen Mary University of London) partnered with the Foreign Office Historians to organize a conference on ‘Contemporary political history in the digital age’. Reflecting on the event, Helen McCarthy noted that ‘Part of the reason historians have been late to the party is because for all but the most contemporary of scholars, the vast bulk of relevant source material is still paper-based’. In other words, it has been possible just to ignore this new, and difficult, material. But she concludes that, with more and more government business being transacted digitally, on the web, ‘historians just have to – well, just go out and start doing it. To grasp the full possibilities and challenges of the digital archive, we need to begin to work with born digital sources’ (McCarthy, 2016). In this instance, it is specifically political history that is being described, and this is where engagement with all kinds of born-digital data is most likely to be seen first. The drift away from the analogue that can be seen in the sphere of government will leave historians with no option but to turn to web archives and other repositories of the digital. In the UK, for example, this is evidenced by a government commitment that ‘In the future our services will be fit for the 21st Century – agile, flexible and *digital by default*’ (my emphasis; Maude, 2013).[[6]](#endnote-6) But while political historians may necessarily have to lead the way, scholars working in other fields can also come to recognize the value of web archives.

The Big UK Domain Data for the Arts and Humanities project (BUDDAH; <http://buddah.projects.history.ac.uk/>)[[7]](#endnote-7) suggests that looking up only briefly from the analogue is all that is required for humanities researchers to be persuaded of the value of web archives. BUDDAH recruited 10 early career researchers from different humanities disciplines who had never encountered web archives before and asked them to develop research projects which captured some of the stories in the archives. Their conclusions were striking: ‘I found that the UK Web Archive has … enormous potential as a research tool for literary researchers … There is a liberating sense, when working within the archive, of exploring both the past and the future, simultaneously – of entering uncharted territory while also rediscovering forgotten artefacts’ (Cran, 2015); ‘Web archives are inordinately rich, rewarding and immense sources of information, but they are also something new and unique. The old methods and mindsets of both historians and archivists will have to be abandoned or at least revised’ (Deswarte, 2015); ‘As our contemporary lives turn more and more to the digital, so too will our historical research, and the BUDDAH project has shown that, with the correct training and the right tools, web archives can be an incredibly useful research resource’ (Taylor, 2015). These researchers, however, had access to a team of experienced scholars, archivists and developers, who could help them to explore this new digital landscape. The real challenge is to communicate the same value to a much wider audience, who will have to navigate the archives largely on their own.

Increasing the number and range of case studies which showcase the richness of information in web archives is an approach which has been shown to work – it clearly helps to address the ‘what’s in it for me?’ question that motivates so many scholars. But what about a more general background awareness of value? It is here that the media, and newspapers in particular, have an important role to play. In recent months, the news cycle has begun to make the case for the significance of web archives. To my knowledge, the first time that web archives found their way into the mainstream media in the UK, in the news rather than technology pages, was in November 2013, when it was reported that the Conservative party had deleted more than a decade’s worth of speeches from its website. The story was given an added news angle because one of those speeches was by the then prime minister David Cameron praising the Internet for ‘making more information available to more people’. It was noted in *The* *Guardian* newspaper that ‘In a remarkable step the party has also blocked access to the Internet Archive's Wayback Machine, a US-based library that captures webpages for future generations, using a software robot that directs search engines not to access the pages’ (Ramesh & Hern, 2015). The need to explain what the Internet Archive was – and this was just three years ago – indicates quite how new this seemed to most people. On this occasion, the British Library was able to say that it had been archiving the Conservative party website since 2004, so the material had been preserved as part of the national historical record. A more frivolous example is the announcement by the BBC in May 2016 that it would be taking down its BBC Food website, removing 11,000 recipes from the public domain. The combination of the BBC and cooking, both staples of British culture, caused uproar and the Internet Archive was again able to come to the rescue. One popular tabloid newspaper, *The Mirror*, reported: the ‘BBC Food website is closing down – here’s how to make sure you don’t lose all your favourite recipes’ and included a link to the relevant pages in the Internet Archive (Curtis, 2016). This time no explanation was needed.

Finally, there have been two very recent examples of web archives in the news. First, in the UK, most of the content on the website of the EU Referendum Leave campaign, including a number of claims which were proven to be inaccurate, was removed and replaced with a page just saying ‘Thank you’. Newspapers again directed everyone to the Internet Archive (<https://web.archive.org/web/20160620174917/http://www.voteleavetakecontrol.org/>); that particular history could not be so easily erased. Then in the US, there was controversy surrounding the taking down of Melania Trump’s website, which contained a claim that she had received a degree in design and architecture from a Slovenian university. Strikingly, when this was covered in *The Independent* newspaper the Internet Archive was introduced as a research resource, with no accompanying explanatory text – a reporter simply noted that the information about Melania Trump’s education ‘has been consistent since at least 2006, according to archives of the site captured by the Internet Archive’ (Revesz, 2016).

Significant in all of these examples has been the requirement for quick and easy access to data that has disappeared from the live web. And it is here that the Internet Archive has a huge advantage over national archives which can only offer restricted access. Journalists are an important user group, especially when it comes to raising awareness, but they are very unlikely to go to a physical library to pursue a story about the web. The rapidly changing news cycle demands a rapid response to emerging stories. Being able to link straight to an archived web page is much more effective than having to include a paragraph of text explaining that a website has been archived but can only be viewed in a handful of libraries – especially when there is a growing expectation of instant, seamless access. Returning to those deleted Conservative party newspapers, the damaging effect of restrictions arising from legal deposit is made explicit in *The Guardian* coverage. While the British Library’s work to archive the site is acknowledged, the report notes that ‘the British Library archive will only be accessible from terminals in its building, raising questions over the Tory commitment to transparency’ (Ramesh & Hern, 2015). There is no doubt that in this case archiving and preservation are not viewed as sufficient – access is key. This is, of course, a journalist’s rather than a historian’s perspective, but historians are not immune to preferring sources which are readily available.

I described the BBC Food incident above as relatively ‘frivolous’, and compared to high politics perhaps it is. But the story gained so much publicity because it affected the daily lives of a large number of people. Much-loved and well-used family recipes were apparently about to be removed from the web forever. And it is as web archives become directly relevant to our social, cultural and working lives that they will truly begin to break in to the mainstream. One area of activity where this is increasingly likely to be the case is the law. In May 2016, for example, a federal judge in the US ruled that the Wayback Machine was a legitimate source of trial evidence, in this instance in a case involving trademark infringement. The main criterion for his decision was that, while the archiving process might miss some data, it did not add or create material that had not originally been present (Bychowski, 2016).[[8]](#endnote-8) There will be more and more examples of this as the web comes to represent the public face of businesses and institutions. How, for example, would a university prove not just what its course handbook contained at a particular date but when and how that information was publicly accessible?

Another topic of debate which is likely to affect public understanding about and awareness of web archives is the growing interest in the individual ‘right to be forgotten’. This is broader than the European Court of Justice ruling against Google in 2014, which (in very simplified terms) allowed individuals within the European Union to request that certain information about them be removed from the search engine’s index (Google Spain SL, Google Inc. v. Agencia Española de Protección de Datos (AEPD), Mario Costeja González, 2014). For most of human history, even for elites, there was very little chance of leaving any kind of trace in the documentary record. Being remembered at all, beyond one or two generations, generally indicated either great achievement or the chance survival of evidence – in essence, luck. The opportunity to be remembered, and the consequent concern about the type of legacy that might be left, is now open to far more people if one discounts the narrative of a ‘digital dark age’ that appears with regularity in the press (another opportunity to demonstrate the value of web archives).[[9]](#endnote-9) Web archives are part of this new landscape, and the relationship between the right of nations to preserve their heritage and the right of individuals to decide how they want to be remembered (or not) will no doubt be negotiated in forums both public and private.

Contested though they may be, web histories are very much here to stay. They are not, however, the whole story when it comes to the record of our digital lives. If historians are already lagging behind in their engagement with the archived web, how much greater will be their difficulty in coming to terms with social media, apps and other forms of born digital data which fall outside the scope of web archives? Some of this information is imperfectly captured by current web harvesting methods, often appearing as a tantalising ghost presence in the archive, an indication of what we do not have. Still more of it is inaccessible or missing altogether, and will never be susceptible to archiving within the frameworks that we have developed for the web. Some of this data is in the hands of businesses, who will only preserve it for as long as it retains commercial value; some is explicitly designed to disappear almost as soon as it is created – ephemerality is built in to the DNA of Snapchat, for example. The voices of individuals, and particularly of the marginalized, may be louder here than they are on the web, so it is all the more important that we look beyond web archives and seek other digital challenges. A web archive does not, and cannot, contain all of our digital data, but it is at least part of the picture. And the relative maturity of web archiving means that there are lessons to be learnt for those seeking to negotiate the complex and evolving digital landscape to piece together 21st-century life.

**Reference List**

Brügger, N., & Finnemann, N. O. (2013). The web and digital humanities: Theoretical and methodological concerns. *Journal of broadcasting and electronic media*, 57, 66-80.

Brügger, N. (2013). Web historiography and internet studies: Challenges and perspectives. *New media & society*, Vol. 15, 752-764.

Bychowski, S. (2016, 16 May). ‘The Internet Archive Wayback Machine: A useful IP litigation tool, but is it admissible?’. Retrieved from <http://www.trademarkandcopyrightlawblog.com/2016/05/the-internet-archive-wayback-machine-a-useful-ip-litigation-tool-but-is-it-admissible/>.

Cran, R. (2015). ‘all writing is in fact cut ups’: The UK Web Archive and Beat literature. Retrieved from <http://sas-space.sas.ac.uk/id/eprint/6101>.

Curtis, S. (2016, 17 May). BBC Food website is closing down – here’s how to make sure you don’t lose all your favourite recipes. *The Independent*. Retrieved from <http://www.mirror.co.uk/tech/bbc-food-website-closing-down-7991786>.

Deswarte, R. (2015). Revealing British Euroscepticism in the UK web domain and archive case study. Retrieved from <http://sas-space.sas.ac.uk/id/eprint/6103>.

Fox, J. (2016, 23 March). ‘The digital dark ages are upon us’. *BloombergView*. Retrieved from <https://www.bloomberg.com/view/articles/2016-03-23/the-digital-dark-ages-are-upon-us>.

Ghosh, P. (2015, 13 February). ‘Googles Vint Cerf warns of “digital Dark Age”’. *BBC Science and Environment*. Retrieved from http://www.bbc.co.uk/news/science-environment-31450389.

Google Spain SL, Google Inc. v. Agencia Española de Protección de Datos (AEPD), Mario Costeja González (2014, 13 May). Retrieved from <http://curia.europa.eu/juris/document/document_print.jsf?doclang=EN&docid=152065>.

Knapton, S. (2015, 11 October). ‘Vital information could be lost in “digital dark age” warns professor’. *The Telegraph*. Retrieved from <http://www.telegraph.co.uk/news/science/science-news/11922192/Vital-information-could-be-lost-in-digital-dark-age-warns-professor.html>.

Maude, F. (2014, 4 December). ‘Foreword’. *Government Digital Inclusion Strategy*. Retrieved from <https://www.gov.uk/government/publications/government-digital-inclusion-strategy/government-digital-inclusion-strategy>.

Maude, F. (2013, 10 December). ‘Foreword’. *Government Digital Strategy: December 2013*. Retrieved from <https://www.gov.uk/government/publications/government-digital-strategy/government-digital-strategy>.

O’Sullivan, D. (2009). *Wikipedia: A new community of practice*. Farnham: Ashgate.

Jackson, A. (2015). *Ten years of the UK Web Archive: What have we saved?* [PowerPoint presentation]. Retrieved from <http://blogs.bl.uk/webarchive/2015/09/ten-years-of-the-uk-web-archive-what-have-we-saved.html>.

McCarthy, H. (2016). Political history in the digital age: The challenges of archiving and analysing born digital sources. *The Impact Blog*. Retrieved from <http://blogs.lse.ac.uk/impactofsocialsciences/2016/03/31/political-history-in-the-digital-age-born-digital-sources/>.

Ramesh, R., & Hern A. (2013, 13 November). Conservative party deletes archive of speeches from internet. *The Guardian*. Retrieved from <https://www.theguardian.com/politics/2013/nov/13/conservative-party-archive-speeches-internet>.

Revesz, R. (2016, 28 July). Melania Trump’s website vanishes from internet after rumours swirl over her university degree. *The Independent*. Retrieved from <http://www.independent.co.uk/news/world/americas/melania-trump-vanishes-biography-website-online-disappears-rumours-plagiarised-speech-education-a7160856.html>.

Taylor, H. (2015). Do online networks exist for the poetry community? Retrieved from <http://sas-space.sas.ac.uk/id/eprint/6105>.

Thomson, L. L. (2011, 1 December). ‘Admissibility of electronic documentation as evidence in U.S. courts’. *Human Rights Electronic Evidence Study*. Retrieved from <http://www.crl.edu/sites/default/files/d6/attachments/pages/Thomson-E-evidence-report.pdf>.

1. I am grateful to the anonymous peer reviewer for this journal for helpful comments and suggestions. [↑](#endnote-ref-1)
2. In the UK, for example, in 2014 it was found that ‘21% of Britain’s population lack the basic digital skills and capabilities required to realise the benefits of the internet. Around a third of small and medium enterprises (SMEs) don’t have a website, and when we include voluntary, community and social enterprises (VCSEs) this figure rises to 50%’ (Maude, 2014). [↑](#endnote-ref-2)
3. Even though he is referring specifically to Wikipedia, O’Sullivan’s description of ‘participatory’ knowledge, a redefining of the public sphere’, is applicable here (O’Sullivan, 2009, p. 10). [↑](#endnote-ref-3)
4. In the digital sphere, terms like ‘archive’ and ‘curation’ have come to be used in ways that would not be recognizable to a professionally trained archivist. Many of our understandings about the creation, purpose and functioning of archives simply do not apply when we are dealing with the ‘archiving’ of the web. The review and transfer procedures that have been developed for paper records, for example, are clearly very different from automated harvesting. There is no paper equivalent to the transformation that occurs when a live web page is converted into an archival format such as WARC or ARC. New and more meaningful vocabularies will undoubtedly develop, but for the moment it is ‘web archive’ that has gained currency. [↑](#endnote-ref-4)
5. The Internet Archive has made a number of datasets available for more complex analysis on a case-by-case basis, for example to the British Library (data from the .uk domain crawled between December 1996 and April 2013; <https://www.webarchive.org.uk/shine>) and to the Alexandria team at the University of Hannover (<http://alexandria-project.eu/>), but this is not scalable; nor can the data be republished. [↑](#endnote-ref-5)
6. There is a noticeable looseness about the terminology used here, with a tendency to treat ‘online’ and ‘digital’ as pretty much synonymous (Maude, 2013). The focus of government strategy is clearly on engaging with its citizens via the web. [↑](#endnote-ref-6)
7. The BUDDAH project, a collaboration between the School of Advanced Study, University of London, Aarhus University, the British Library and the Oxford Internet Institute, was funded by the Arts and Humanities Research Council (Award no. AH/L009854/1). [↑](#endnote-ref-7)
8. Indeed, Thomson notes that, as early as 2004, ‘Archived information from websites is being admitted into evidence just as information from active sites is being admitted’ (Thomson, p. 25 and n. 49). [↑](#endnote-ref-8)
9. There are far too many examples of this digital doom-mongering to list them all, but recent examples include Fox (2016), Ghosh (2015) and Knapton (2015). [↑](#endnote-ref-9)